International Journal of Zoology
and Research (IJZR)
ISSN(P): 2278-8816; ISSN(E): 2278-8824
Vol. 6, Issue 6, Dec 2016, 1-12
© TJPRC Pvt. Ltd

TRANS STELLAR
•Journal Publications •Research Consultancy

HEMICHORDATE EVOLUTIONARY RELATIONSHIPS BASED ON CODING GENES

BIBHUTI PRASAD BARIK

Assistant Professor, Department of Zoology, Khallikote University, Brahmapur, Odisha, India

ABSTRACT

Evolutionary relationship within hemichordates is very interesting. The objectives of the current study were to infer evolutionary relationship among hemichordates based on coding genes (ATPase alpha, cytochrome b and histone) along with in silico proteomic analysis. Evolutionary relationships were inferred showing similar species clustered together but did not form distinct clades as per their lineage and morphological similarities. It was noticed that some species appeared to be polyphyletic and this could be assumed by possible mutations and adaptive radiations. The variation in nucleic acid composition is observed and may be attributed to mutational pressure. GC content of the genes was predicted and significantly varied. This variation might have played a crucial role in lineage based on patterns of base composition within and among species. The amino acid composition of the translated proteins, atomic composition of corresponding amino acids and physiochemical parameters has been used to determine the evolutionary trend among hemichordates. The species have evolved and expected to have adapted to different ecological environment.

KEYWORDS: Hemichordate, Coding Genes, Evolution, Translation, In Silico & Physiochemical Properties

Original Article

Received: Oct 03, 2016; Accepted: Oct 24, 2016; Published: Oct 28, 2016; Paper Id.: IJZRDEC20161

INTRODUCTION

Evolutionary studies of hemichordates are significant in understanding chordate evolution. Molecular phylogenetic analyses only have not yet provided robust support for the hemichordate evolution. The objectives of the current study were to infer phylogenetic relationship among hemichordates based on ATPase alpha, cytochrome b and histone genes along with *in silico* proteomic analysis of the translated proteins.

MATERIALS AND METHODS

Retrieval of Sequences and Taxon Sampling

The selected coding gene sequences of hemichordates available in Gen Bank (Benson et al., 2013) database were retrieved using a PERL script. The sequences were filter searched and were selected based on gene types using Bioedit software version 7.0.5.3 (Hall, 1999). These gene sequences were considered for phylogenetic analysis and their corresponding proteomic analysis were also carried out.

Multiple Sequence Alignment and Phylogenetic Analysis

The retrieved gene sequences were saved and fasta formatted for multiple sequence alignment. The sequences were aligned using CLUSTAL W (Thomption et al., 1994). For pair wise sequence alignment the gap opening penalty and extension penalties were 15 and 6.66 respectively. The aligned file was exported for phylogenetic analysis. Five different methods (ML, NJ, ME, UPGMA and MP) were adopted to perform phylogenetic analysis using MEGA 7 software (Kumar et al., 2016). The branch length and consistency, retention and composites indices are shown in table 1.

Table 1: Branch length and indices of CI, RI and CI

Sl.No.	Gene	Sum of Branch Length				Consistency Index	Retention Index	Composite Index
		ML	NJ	ME	UPGMA			
1.	ATPase alpha	-501	2.166	2.166	2.211	0.837	0.806	0.781
2.	Cytochrome b	-156	1.076	1.076	1.081	0.685	0.453	0.340
3.	Histone	-360	9.455	9.455	9.098	0.458	0.887	0.411

ML: Maximum Likelihood, **NJ:** Neighbour Joining, **ME:** Minimum Evolution, **UPGMA:** Unweighted Pair Group Method with Arithmetic Mean**, MP:** Maximum Parsimony, **CI:** Consistency Index, **RI:** Retention Index and **CI:** Composite Index.

All characters were equally weighted and unordered. Alignment gaps were treated as missing data. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap was 500 replicates. The evolutionary distances were computed.

Nucleic Acid Composition

Nucleotide composition of the gene sequences were computed using Bioedit (Hall, 1999).

Translation and *In silico* Physiochemical Characterization

The gene sequences were translated to their corresponding protein sequences and the compositions of amino acids were predicted using Bioedit. The physicochemical properties such as atomic composition, molecular weight, theoretical pI, instability indices, aliphatic indices and grand average of hydropathicity etc. were computed using ExPASy's ProtParam tool (Gasteiger et al., 2005).

RESULTS & DISCUSSIONS

Phylogenetic Analysis

Maximum Likelihood Trees

The evolutionary history was inferred by using the Maximum Likelihood method based on the Tamura-Nei model (Tamura and Nei, 1993). The trees were with the highest log-likelihood. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pair wise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated (Figure 1-3).

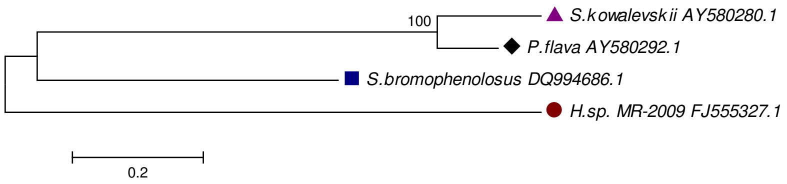
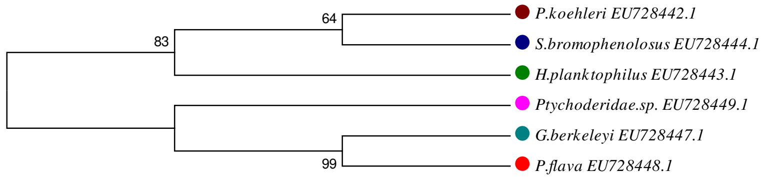

100
▲ S.kowalevskii AY580280.1
◆ P.flava AY580292.1
■ S.bromophenolosus DQ994686.1
● H.sp. MR-2009 FJ555327.1
0.2

Figure 1: ML Tree Based on ATPase Alpha Gene



64
83
99
● P.koehleri EU728442.1
● S.bromophenolosus EU728444.1
● H.planktophilus EU728443.1
● Ptychoderidae.sp. EU728449.1
● G.berkeleyi EU728447.1
● P.flava EU728448.1

Figure 2: ML Tree based on Cytochrome b Gene

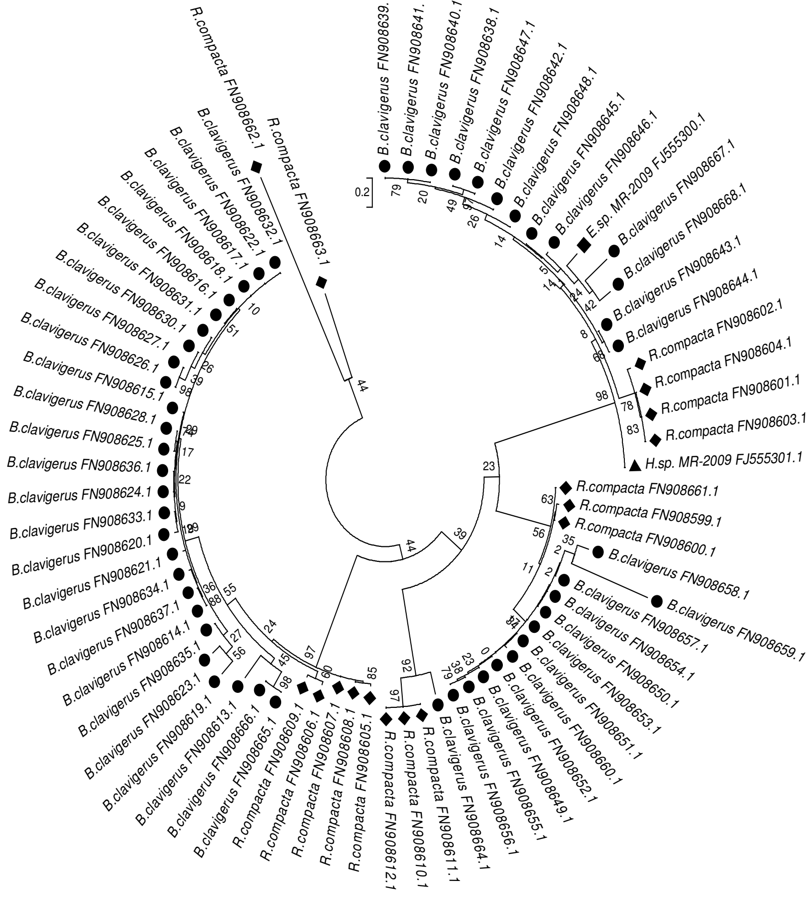
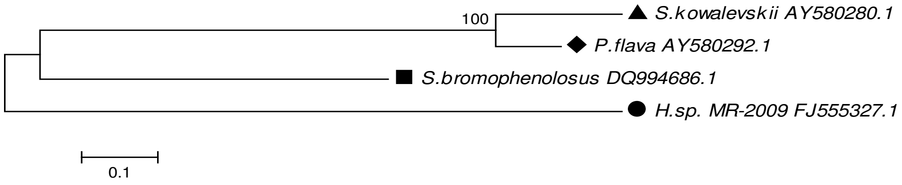


Figure 3: ML Tree based on Histone Gene

Neighbor Joining Trees

The evolutionary history was inferred using the Neighbor-Joining method (Saitou and Nei, 1987). The optimal trees were drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic trees (Figure 4-6). The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura et al., 2004) and are in the units of the number of base substitutions per site.



100
▲ S.kowalevskii AY580280.1
◆ P.flava AY580292.1
■ S.bromophenolosus DQ994686.1
● H.sp. MR-2009 FJ555327.1
0.1

Figure 4: NJ Tree based on ATPase Alpha Gene

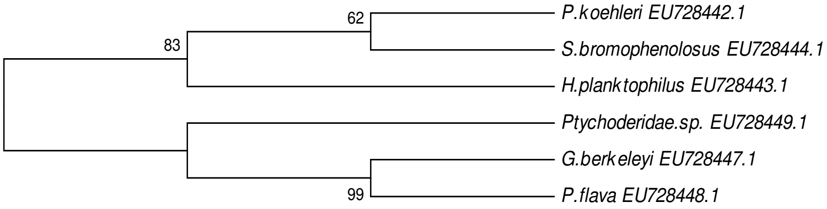

62
83
99
P.koehleri EU728442.1
S.bromophenolosus EU728444.1
H.planktophilus EU728443.1
Ptychoderidae.sp. EU728449.1
G.berkeleyi EU728447.1
P.flava EU728448.1

Figure 5: NJ Tree based on Cytochrome b Gene

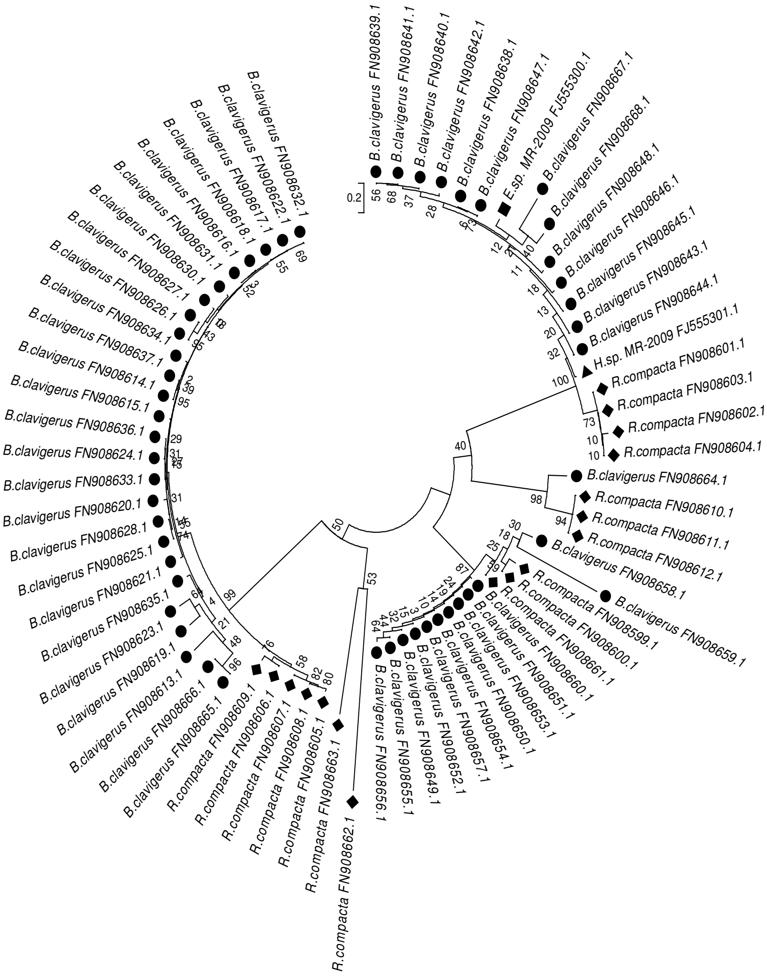


Figure 6: NJ Tree based on Histone Gene

Minimum Evolution Trees

The evolutionary history was inferred using the Minimum Evolution method (Rzhetsky and Nei, 1992). The trees are drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic trees (Figure 7-9). The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura et al., 2004) and were in the units of the number of base substitutions per site. The ME trees were searched using the Close-Neighbor-Interchange (CNI) algorithm (Nei and Kumar, 2000).

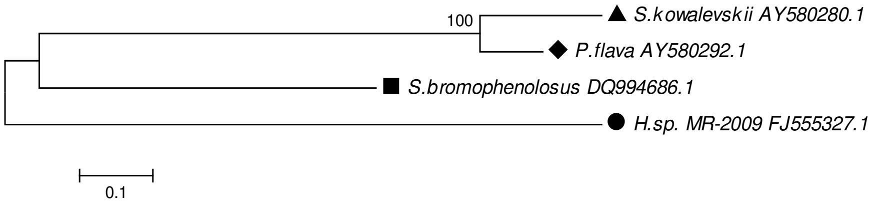

100
▲ S.kowalevskii AY580280.1
◆ P.flava AY580292.1
■ S.bromophenolosus DQ994686.1
● H.sp. MR-2009 FJ555327.1
0.1

Figure 7: ME Tree based on ATPase Alpha Gene

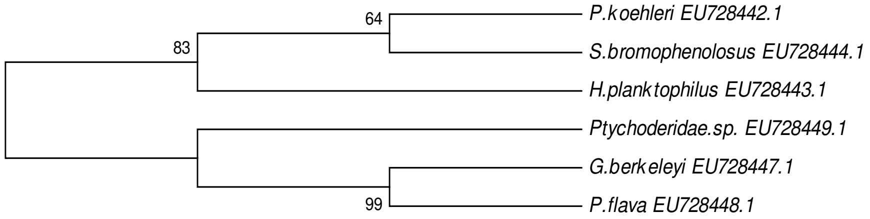

64
83
P.koehleri EU728442.1
S.bromophenolosus EU728444.1
H.planktophilus EU728443.1
Ptychoderidae.sp. EU728449.1
G.berkeleyi EU728447.1
99
P.flava EU728448.1

Figure 8: ME Tree Based on Cytochrome b Gene

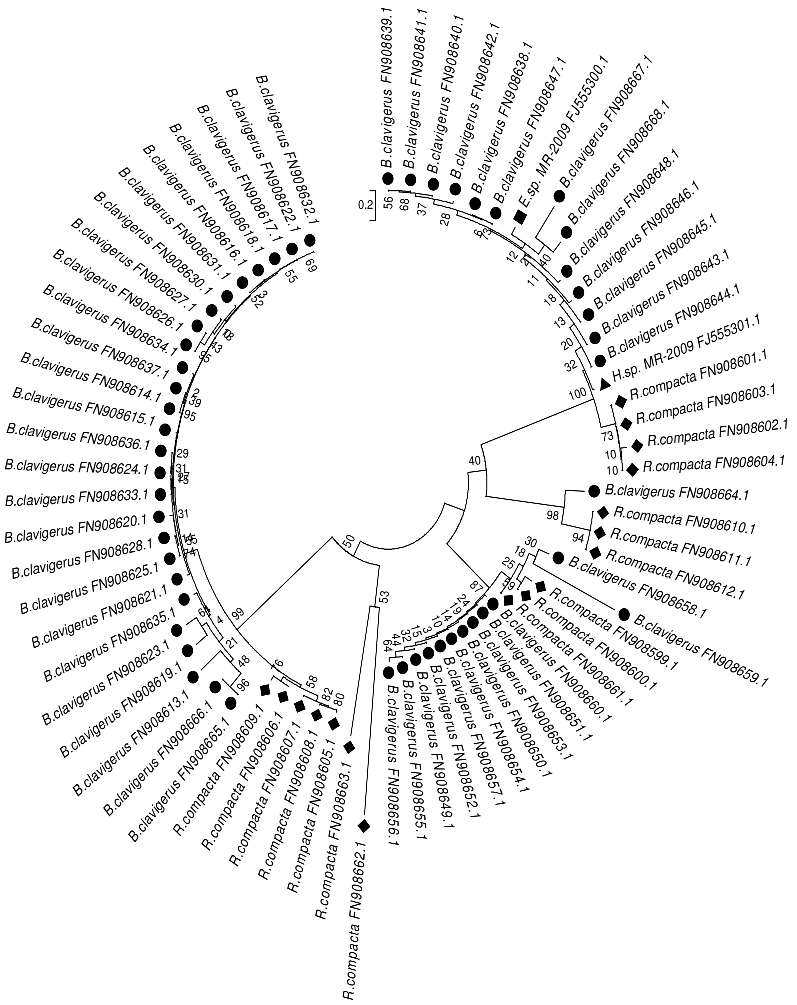


Figure 9: ME Tree based on Histone Gene

UPGMA Trees

The evolutionary history was inferred using the UPGMA method (Sneath and Sokal, 1973). The optimal trees were drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic trees (Figure 10-12). The evolutionary distances were computed using the Maximum Composite Likelihood method and were in the units of the number of base substitutions per site.

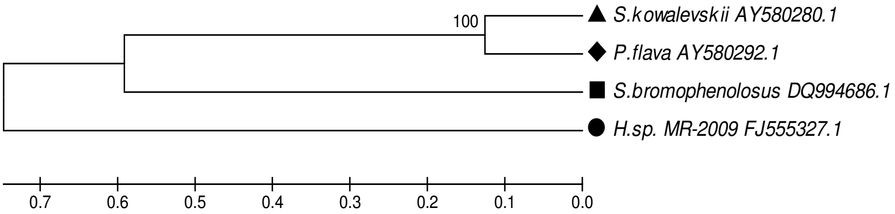

100
▲ S.kowalevskii AY580280.1
◆ P.flava AY580292.1
■ S.bromophenolosus DQ994686.1
● H.sp. MR-2009 FJ555327.1
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

Figure 10: UPGMA Tree Based on ATPase Alpha Gene

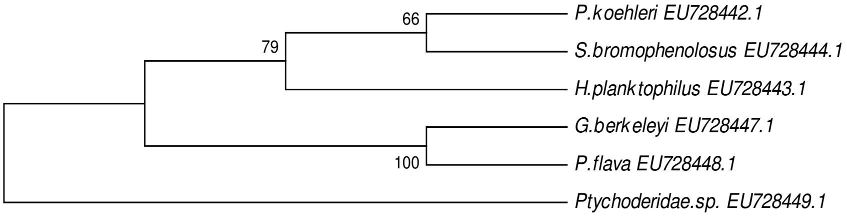

66
79
P.koehleri EU728442.1
S.bromophenolosus EU728444.1
H.planktophilus EU728443.1
G.berkeleyi EU728447.1
100
P.flava EU728448.1
Ptychoderidae.sp. EU728449.1

Figure 11: UPGMA Tree Based on Cytochrome b Gene

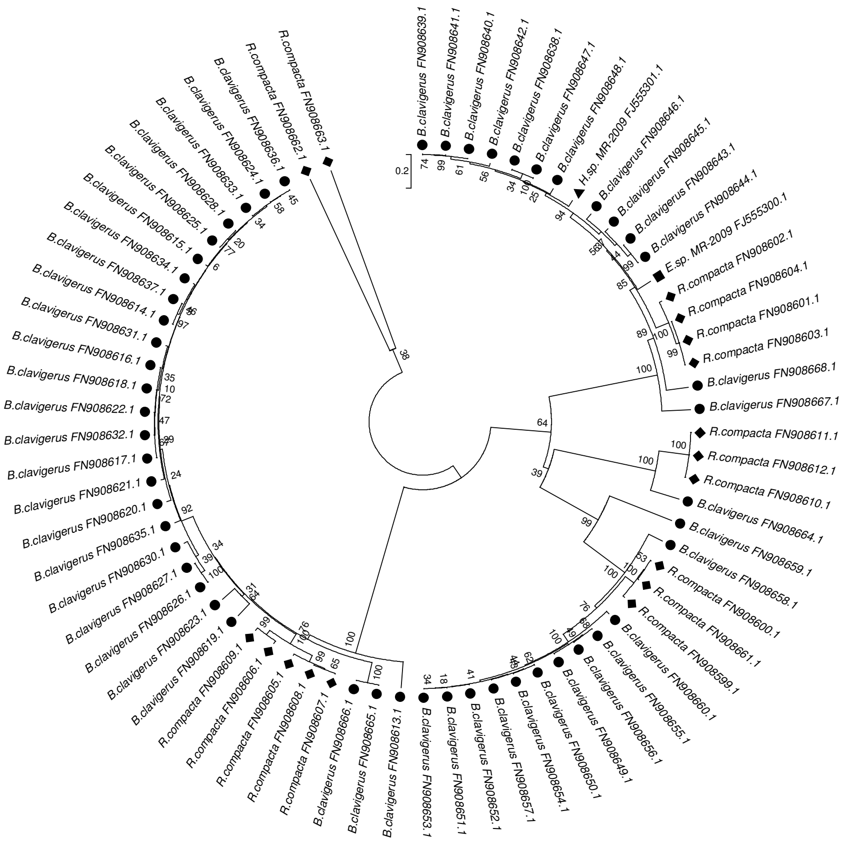


Figure 12: UPGMA Tree based on Histone Gene

MP Trees

The evolutionary history was inferred using the Maximum Parsimony method (Figure 13-15). The MP trees were obtained using the Subtree-Pruning-Regrafting (SPR) algorithm with search level 0 in which the initial trees were obtained by the random addition of sequences (10 replicates). All positions containing gaps and missing data were eliminated.

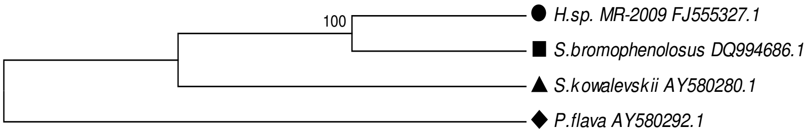

100
● H.sp. MR-2009 FJ555327.1
■ S.bromophenolosus DQ994686.1
▲ S.kowalevskii AY580280.1
◆ P.flava AY580292.1

Figure 13: MP Tree based on ATPase Alpha Gene

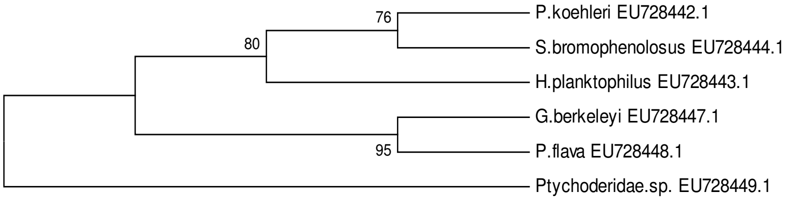

76
80
95
P.koehleri EU728442.1
S.bromophenolosus EU728444.1
H.planktophilus EU728443.1
G.berkeleyi EU728447.1
P.flava EU728448.1
Ptychoderidae.sp. EU728449.1

Figure 14: MP Tree based on Cytochrome b Gene

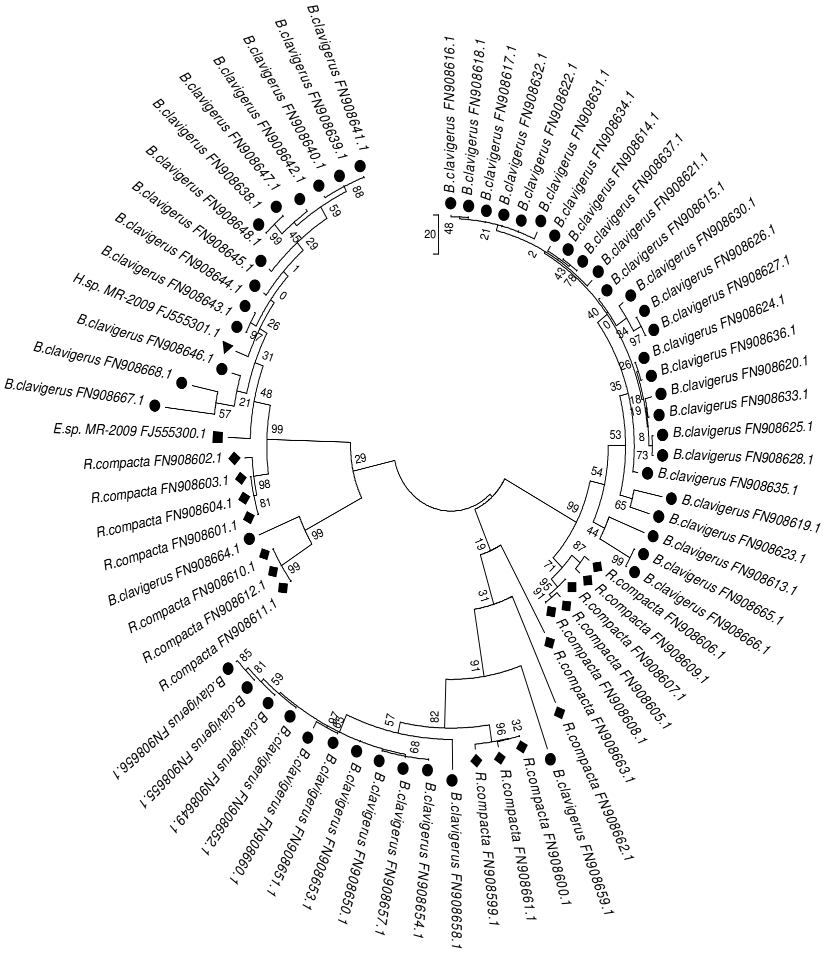


Figure 15: MP Tree based on Histone Gene

Recently Osborn and colleagues showed that despite their great morphological diversity, most deep-living hemichordates form a single clade (Osborn et al., 2012). In the current study, ATPase alpha, cytochrome b and histone genes sequences were used to infer for phylogenetic affiliations among species belonging to different species of hemichordates. Phylogenetic trees were investigated by different methods including most parsimonious trees to infer phylogeny. The tree showed more or less similar species clustered together but did not form distinct clades as per their lifestyles. The result also indicated that several species appear to be polyphyletic and several unrelated species appear to share the same clade.

Nucleic Acid Composition

Nucleotide composition of the gene sequences were predicted (Table 2) and the result showed that GC content of ATPase alpha gene was below 50% i.e. 47.53% in case of *P. flava* and lowest (42.52%) in case of *S. bromophenolosus*. GC content of cytochrome b gene was high i.e. 62.9% in case of *Ptychoderidae. sp* and lowest (53%) in case of *P. flava.* Similarly in case of histone gene *Enteropneusta.sp. contains* 45.65% of GC with lowest (8.11%) in *R. compacta.* The variation in nucleic acid composition is observed and may be attributed to mutational pressure. In the representative species of hemichordata, i.e., Balanoglossus carnosus, possible influence of mutational pressure due to compositional constraints in codon usage was noticed (Karumathil, 2016).

Table 2: Nucleic Acid Composition of Gene Sequences of Different Species

Gene	Species	Sequence Length	A+T %	G+C %	A	T	G	C	Molecular Weight

ATPase alpha	*Hemichordata. sp.*	1208	52.98	47.02	315	325	307	261	26.08(A)
									21.61(C)
									25.41(G)
									26.90(T)

	S. kowalevskii	1233	56.53	43.47	375	322	297	239	30.41(A)
									19.38(C)
									24.09(G)
									26.12(T)

	P. flava	1233	52.47	47.53	338	309	311	275	27.41(A)
									22.30(C)
									25.22(G)
									25.06(T)

	S. bromophenolosus	1103	57.48	42.52	321	313	261	208	29.10 (A)
									18.86(C)
									23.66(G)
									28.38(T)

Cytochrome b	*P. koehleri*	421	55.34	44.66	100	133	50	138	23.75(A)
									32.78(C)
									11.88(G)
									31.59(T)

	H. planktophilus	306	54.25	45.75	72	94	30	110	23.53(A)
									35.95(C)
									9.80(G)
									30.72(T)

	S. bromophenolosus	386	58.81	41.19	93	134	50	109	24.09(A)
									28.24(C)
									12.95(G)
									34.72(T)

	G. berkeleyi	392	54.08	45.9	101	111	55	125	25.77(A)
									31,89(C)
									14.03(G)
									28.32(T)

	P. flava	405	53.0	46.9	115	100	53	137	28.40 (A) 33.83 (C) 13.09(G)

									24.69(T)

	Ptychoderidae. sp.	386	62.9	37.05	116	127	40	103	30.05(A)
									26.68(C)
									10.36(G)
									32.90(T)

Histone	*B. clavigerus*	1074	4.28	6.05	23	23	39	26	2.14%(A) 2.42%(C) 3.63%(G) 2.14%(T)
	Enteropneusta.sp.	276	45.65	54.35	75	51	72	78	27.17%(A) 28.26%(C) 26.09%(G) 18.48%(T)
	Hemichordata. sp.	276	44.2	53.26	61	61	72	75	22.10%(A) 27.17%(C) 26.09%(G) 22.10%(T)
	R. compacta	456	8.11	14.91	15	22	36	32	3.29%(A) 7.02%(C) 7.89%(G) 4.82%(T)

GC content of the genomic DNA is significantly varied and this variation might have played a significant role in the evolution. Two main evolutionary processes have been raised to explain the patterns of variation of base composition within and among species: biases in the process of mutation, such that the rates of change from G-C and A-T are not constant in time or space (Sueoka, 1988); and natural selection, either on overall GC content or on specific patterns of codon usage (Eyre-Walker, 1999). The most active neutralist–selectionist debate has concern of GC evolution in vertebrates (Mooers and Holmes, 2000). Such explicitly phylogenetic studies of GC dynamics are rare, perhaps because of concerns that changes in base composition can affect the accuracy with which we reconstruct trees. The effects of GC variation on phylogenetic inference need to be explored further. The phylogenetic effects of GC differences apply to neighbouring nucleotide sites (Karlin and Mrazek, 1997).

Translation and *In silico* Physiochemical Characterization

The amino acid composition of proteins has been used to determine the trend among species (Bogatyreva, 2006) more specifically hemichordates to understand its evolutionary perspective (Sorimachi, 1999) and to identify the contrasting feature of proteins (Gaur, 2010). The species have evolved and adapted to different ecological environment. Their common origin indicates that amino acid composition of proteins of different kinds of eukaryotes may be similar. However, it is interesting to observe the contrasting features introduced due to their local ecological adjustment (Table 3).

Table 3: Amino Acid Composition of Translated Proteins of Different *Species*

Protein	Species	Molecular Weight	Maximum no. Amino Acids
ATPase alpha	*Hemichordata.sp.*	99696.46	Thr 325
	S.kowalevskii	100817.78	Ala 375
	P.flava	101385.63	Ala 338
	S.bromophenolosus	90818.96	Ala 321
Cytochrome b	*P.koehleri*	37657.14	Cys 138
	H.planktophilus	27695.36	Cys 110
	S.bromophenolosus	34269.69	Thr 134
	G.berkeleyi	34448.44	Cys 125
	P.flava	35454.90	Cys 137
	Ptychoderidae.sp.	34007.26	Thr 127
	P.bahamensis	46433.81	Ala 153
	H.kupfferi	45275.08	Cys 161
	M.psammophilus	47181.50	Cys 174
Histone	*B.clavigerus*	118762.82	Asn 963
	Enteropneusta.sp.	22656.89	Cys 78
	Hemichordata.sp.	23456.61	Cys 75
	R.compacta	48711.65	Asn 351

The percentage of occurrence of the amino acids in proteins depends, at least for some of the residues, on the protein dimension (Carugo, 2008). The amino acid composition along with atomic composition (Table 4) of conserved residues in present-day proteins, i.e., those residues which are unchanged between an ancestral sequence and any given descendant sequence, is determined by two factors the amino acid composition within ancestral sequences, and the relative probability of conservation of each amino acid between an ancestral and an extant descendant sequence (Brooks et al., 2002).

Table 4: Atomic Composition (No. of Carbon, Hydrogen, Nitrogen, Oxygen and Sulphur) of Protein Sequences of Different Species

Protein	Species name	Carbon	Hydrogen	Nitrogen	Oxygen	Sulphur
ATPase alpha	*Hemichordata.sp.*	1986	3261	547	524	28
	P.flava	2024	3213	547	597	16
	S.bromophenolosus	1755	2802	462	549	12
	S.kowalevskii	2009	3212	552	586	18
Cytochrome b	*G.berkeleyi*	690	1014	156	172	1
	H.planktophilus	550	827	153	133	1
	P.flava	701	1157	233	171	2
	P.koehleri	746	1103	165	189	1
	Ptychoderidae.sp	683	1014	150	175	0
	S.bromophenolosus	678	1000	150	173	1
Histone	*B.clavigerus*	190	300	60	52	1
	Enteropneusta sp	459	766	142	126	1
	R.compacta	177	328	86	49	0

Physiochemical properties of the translated proteins were computed (Table 5). The molecular weight of the proteins ranged from 45275 to 4288.9. The computed isoelectric point (pI value) of all species ranges from 12.48 to 4.96. The instability indices indicated higher values in *R.compacta* and those of *Ptychoderidae.sp* was lower. The aliphatic index refers to the relative volume of a protein that is occupied by aliphatic side chains and contributes to the increased thermostability of protein. In the present study aliphatic indices of *Ptychoderidae sp.* (116.37) were found to be higher than those of others. This indicates that proteins of this species are more stable than those of others over a wide temperature range. Thus it may be assumed that *Ptychoderidae* species are more likely to change and adapt to varied environments. Grand average of hydropathicity (GRAVY) values indicates the solubility of proteins: negative GRAVY values of most hemichordate species showed it to be hydrophilic in nature with few exceptions indicating a little surface accessibility of the protein to interact with water.

Table 5: *In silico* Physiochemical Properties of Proteins of Different Species of Hemichordates

Protein	Species	MW	Tpl	II	AI	Gravy
ATPase alpha	*Hemichordata.sp.*	44083.7	10.71	61.89	111.75	0.194
	P.flava	45275.0	8.61	35.89	80.39	-0.353
	S.bromophenolosus	39543.0	4.96	26.09	93.32	-0.155
	S.kowalevskii	45052.0	9.04	29.78	82.34	-0.291
Cytochrome b	*G.berkeleyi*	14278.6	6.12	28.28	111.26	0.709
	H.planktophilus	11742.6	11.77	87.33	70.91	-0.600
	P.flava	15649.4	12.43	73.92	81.97	-0.966
	P.koehleri	15439.0	5.25	29.85	117.39	0.746
	Ptychoderidae.sp	14126.4	5.27	25.49	116.37	0.600
	S.bromophenolosus	14052.3	5.23	35.28	115.32	0.698
Histone	*B.clavigerus*	4288.9	10.95	73.12	73.78	-0.511
	Enteropneusta sp	10322.0	11.02	49.94	83.91	-0.551
	Hemichordata sp	10561.7	11.02	40.91	73.15	-0.710
	R.compacta	4445.1	12.48	192.55	55.71	-2.151

M.W: Molecular Weight, T.pI: Theoretical pI, I.I: Instability Index, A.I: Aliphatic Index, GRAVY: Grand Average of Hydropathicity.

CONCLUSIONS

The accumulation of mutations can eventually lead to differences between ATPase alpha, cytochrome b and histone proteins with respect amino acid composition (Table 5). Thus, over a very long evolutionary time, mutation and drift (Hughes, 2010) appear to be able to overcome the conservative effect of stabilizing selection on physiochemical properties of proteins and give rise to a certain degree of functional differentiation.

REFERENCES

1. *Benson D.A., Cavanaugh M., Clark K., Karsch-Mizrachi I., Lipman D.J., Ostell J. and Sayers E.W. (2013) GenBank. Nucleic Acids Res 41: 36-42.*
2. *Bogatyreva N.S., Finkelstein A.V., Galzitskaya O.V. (2006) Trend of amino acid composition of proteins of different taxa. J Bioinfor Comput Biol 4(2):597-608.*
3. *Brooks D.J., Fresco J.R., Lesk A.M. and Sing M. (2002) Evolution of Amino Acid Frequencies in Proteins Over Deep Time: Inferred Order of Introduction of Amino Acids into the Genetic Cod Mol. Biol. Evol 19(10):1645-1655.*
4. *Carugo O. (2008) Amino acid composition and protein dimension, Protein Science 17:2187-2191.*
5. *Eyre-Walker A. (1999) Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. Genetics 152: 675-683.*
6. *Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R.D., and Bairoch, A. (2005) Protein identification and analysis tools on the ExPASy Server. In J. M. Walker (Ed.), The Proteomics Protocols Handbook (pp. 571-607) Humana Press.*

7. *Gaur R.K., Natekar G. (2010) Prokaryotic and eukaryotic integral membrane proteins have similar architecture. Mol Biol Rep 37(3):1247-1251.*

8. *Hall T.A. (1999) Bioedit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nuc Acids Symp Ser 41:95-98.*

9. *Hughes A.L. (2010) Evolutionary Conservation of Amino Acid Composition in Paralogous Insect Vitellogenins, Gene 467(1-2): 35-40.*

10. *Karlin S. and Mrazek J. (1997) Compositional differences within and between eukaryotic genomes. Proc Natl Acad Sci USA. 94:10227-10232.*

11. *Karumathil S., Dirisala V.R., Srinadh U., Nikhil V., Kumar N.S.S. and Nair R.R. (2016) Evolution of Synonymous Codon Usage in the Mitogenomes of Certain Species of Bilaterian Lineage with Special Reference to Chaetognatha. Bioinform Biol Insights 10: 167-184.*

12. *Kumar S., Stecher G. and Tamura K. (2016) MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol. Biol. Evol 33(7):1870-1874.*

13. *Mooers, A.O. and Holmes E.C. (2000) The evolution of base composition and phylogenetic inference. Trends Ecol. Evol 15:365-369.*

14. *Nei M. and Kumar S. (2000) Molecular Evolution and Phylogenetics. Oxford University Press, New York.*

15. *Osborn K.J., Kuhnz L.A., Priede I.G., Urata V, Gebruk A.V. and Holland N.D. (2012) Diversification of acorn worms (Hemichordata, Enteropneusta) revealed in the deep sea. Proc. R. Soc. B. 279: 1646-1654.*

16. *Rzhetsky A. and Nei M. (1992) A simple method for estimating and testing minimum evolution trees. Molecular Biology and Evolution 9: 945-967.*

17. *Saitou N. and Nei M. (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4: 406-425.*

18. *Sneath P.H.A. and Sokal R.R. (1973) Numerical Taxonomy. Freeman, San Francisco.*

19. *Sorimachi, K. (1999) Evolutionary changes reflected by the cellular amino acid composition. Amino Acids 17(2), 207-226.*

20. *Sueoka N. (1988) Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA. 85(8):2653-2657.*

21. *Tamura K. and Nei M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Molecular Biology and Evolution 10: 512-526.*

22. *Tamura K., Nei M. and Kumar S. (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. Proceedings of the National Academy of Sciences (USA) 101:11030-11035.*

23. *Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucl. Acids Res 22: 4673-4680*